

Parallelism and Power in the Age of Petascale Computing

Horst D. Simon and John Shalf

Lawrence Berkeley National Laboratory

ACTS Workshop
Berkeley, CA
August 23, 2007



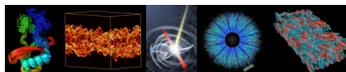
Overview

News from NERSC

My (TOP500) Predictions for Petascale Computing

Three Challenges for Petascale Computing:

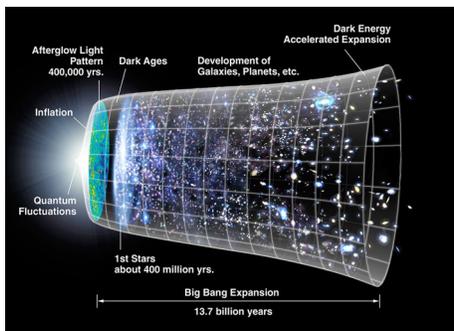
- Parallelism
- Power
- The Petascale bubble



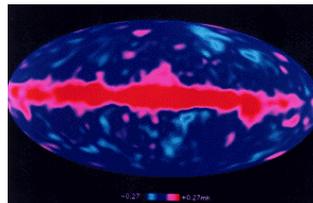
NERSC 5 arrives, January 16, 2007



NERSC User George Smoot wins 2006 Nobel Prize in Physics



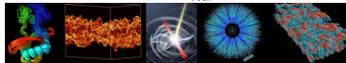
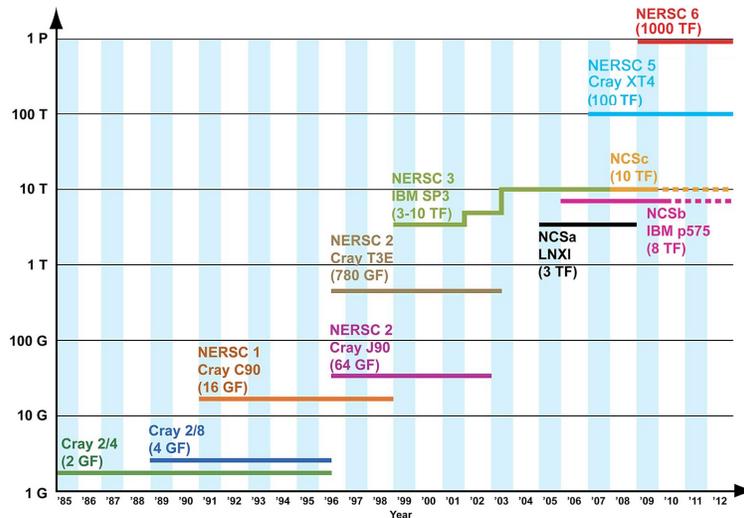
Smoot and Mather 1992
COBE Experiment showed
anisotropy of CMB



**Cosmic Microwave
Background Radiation
(CMB): an image of the
universe at 400,000 years**



NERSC Systems History



Overview

News from NERSC

My (TOP500) Predictions for Petascale Computing

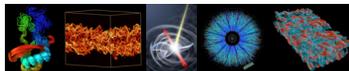
Three Challenges for Petascale Computing:

- Scaling
- Power
- The Petascale bubble

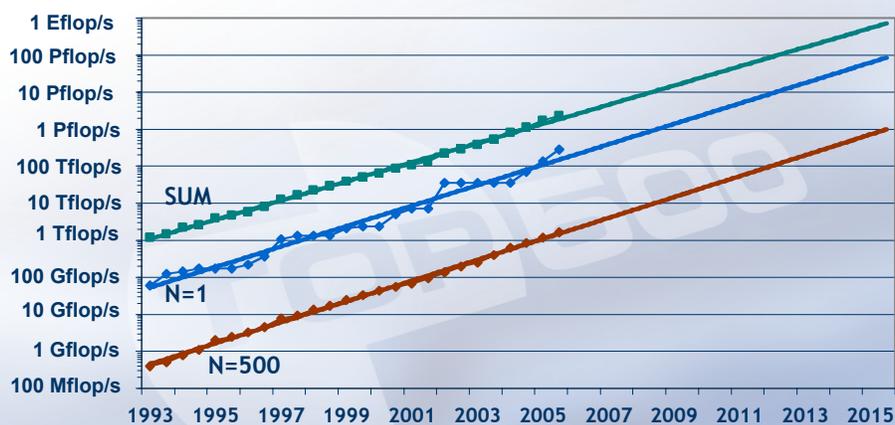


Levels of Petascale Computing

- The term “Petascale” is frequently used, but unfortunately ill-defined
- We need to distinguish
 - Theoretical peak petaflop/s systems
 - LINPACK Rmax Petaflop/s systems (used in TOP500)
 - Sustained applications performance in excess of a Petaflop/s
- My Definition: “**Petascale Computing**”
 - Widespread use of systems that deliver sustained applications performance a level above 1 Petaflop/s
 - Reached when all system on the TOP500 list have more than 1 Petaflop/s Rmax performance

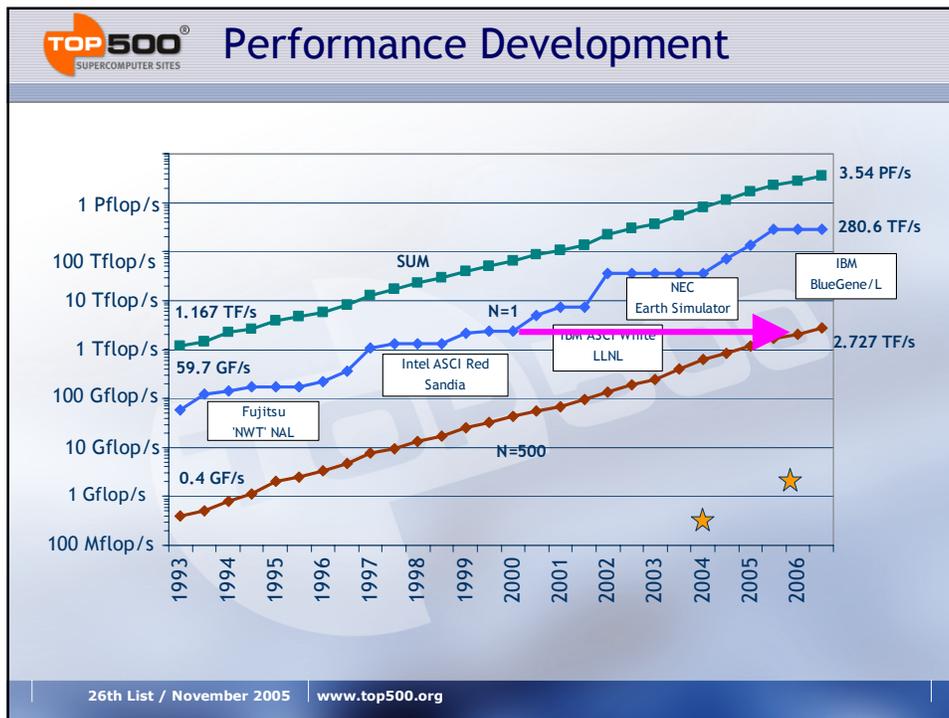
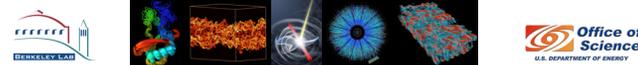


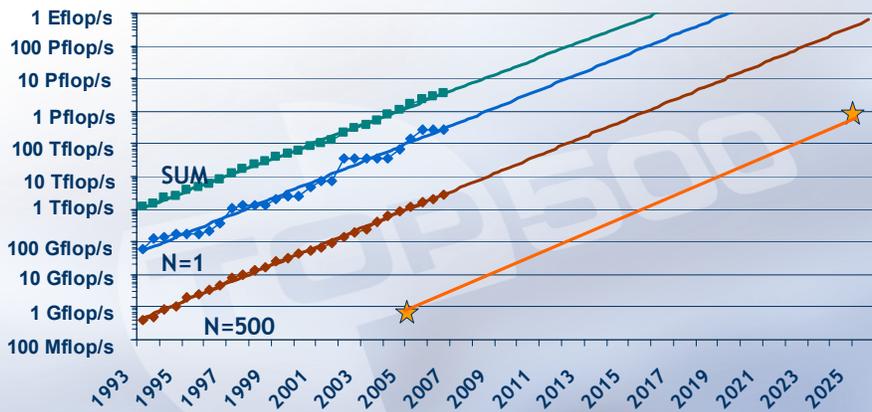
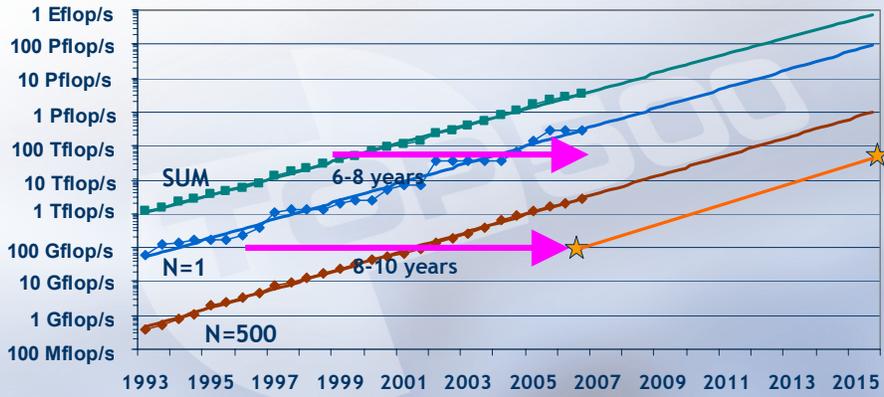
Performance Projection



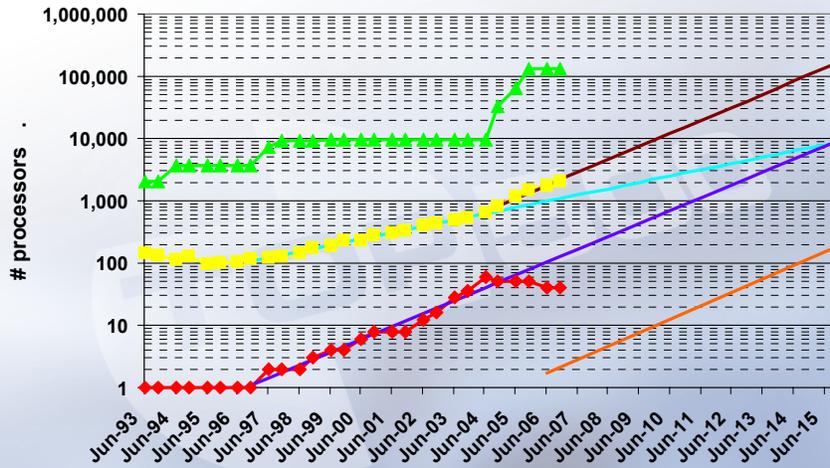
TOP500 Projections

- June 1997:
 - First LINPACK Teraflop/s system tops the list
- June 2005 (8 years later): Terascale computing arrives
 - 1 Teraflop/s is required to enter the TOP500 list
- November 2008:
 - First LINPACK Petaflop/s system tops the list
- June 2016 (7.5 years later): Petascale computing arrives
 - 1 Petaflop/s is required to enter the TOP500 list
- November 2018:
 - First LINPACK Exaflop/s

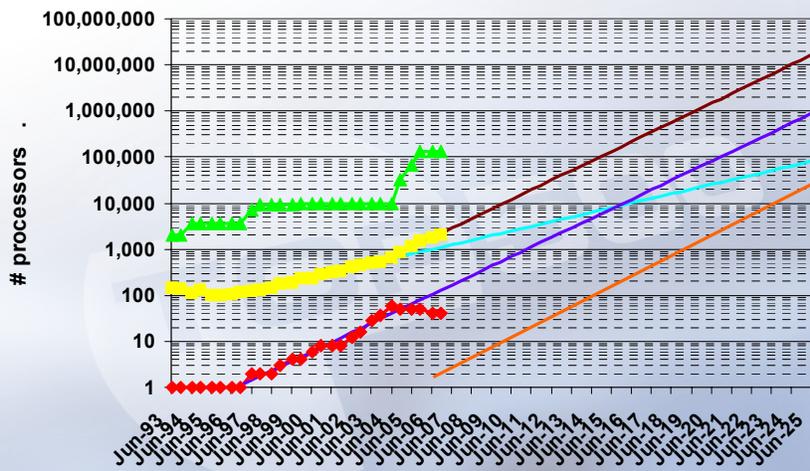




Concurrency Levels



Concurrency Levels- there is a massively parallel system also in your future



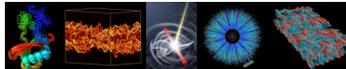
Overview

News from NERSC

My (TOP500) Predictions for Petascale Computing

Three Challenges for Petascale Computing:

- **Parallelism**
- Power
- The Petascale bubble



Traditional Sources of Performance Improvement are Flat-Lining

- New Constraints
 - 15 years of *exponential* clock rate growth has ended
- But Moore's Law continues!
 - How do we use all of those transistors to keep performance increasing at historical rates?
 - Industry Response: #cores per chip doubles every 18 months *instead* of clock frequency!
- Is multicore the correct response?

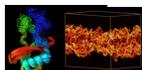
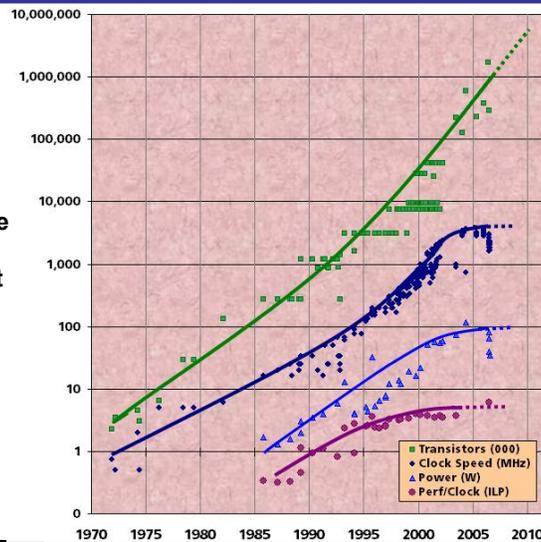
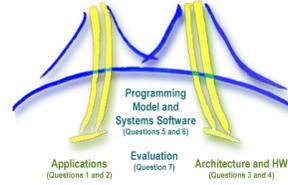


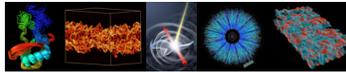
Figure courtesy of Kunle Olukotun, Lance Hammond, Herb Sutter, and Burton Smith

Is Multicore the Correct Response?

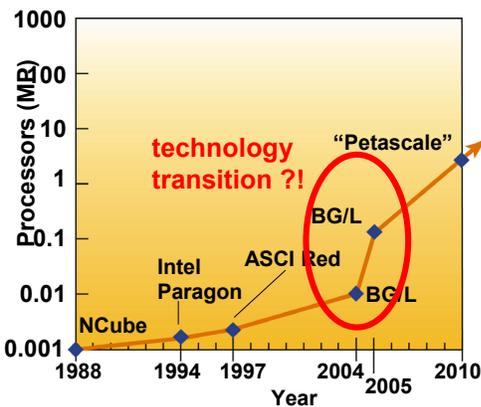
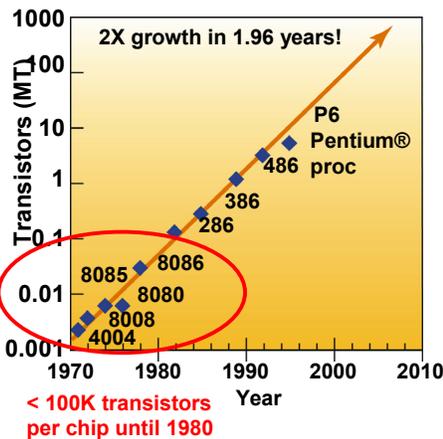
- “The View from Berkeley”,
http://view.eecs.berkeley.edu/wiki/Main_Page



- **Kurt Keutzer:** “This shift toward increasing parallelism is not a triumphant stride forward based on breakthroughs in novel software and architectures for parallelism; instead, this plunge into parallelism is actually a retreat from even greater challenges that thwart efficient silicon implementation of traditional uniprocessor architectures.”
- **David Patterson:** “Industry has already thrown the hail-mary pass. . . But nobody is running yet.”



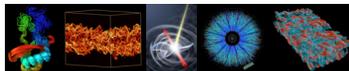
Historical Reference: Transistor Count



“The Processor is the new Transistor”

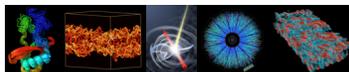
(David Patterson)

- NERSC’s computing system, Seaborg, contains as many processors as there are transistors in the original Intel 8080a implementation (6,000 transistors vs 6,000 processors)
- BG/L at LLNL contains as many processors as there are transistors in the MC68000 (manufactured in 1980, the MC68000L was a 32-bit processor and contained 68,000 transistors).
- With 1.5M processors, BG/Q likely to have more processors than there are logic gates in its constituent processing elements. (is that ironic or is it outrageous?)



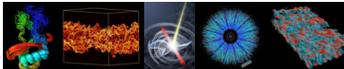
Three Common Misperceptions about Multicore

- Multicore is just like SMP on a chip
- Multicore will be just like programming a very large parallel system in MPI
- Multicore will make the “memory wall” problem worse



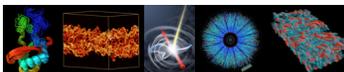
Multicore is NOT an SMP-on-a-Chip

- What about SMP on a chip?
 - Hybrid Model: *Long and mostly unsuccessful history*
 - But multicore is NOT an SMP on a chip
 - 10-100x higher bandwidth on chip
 - 10-100x lower latency on chip
 - SMP model ignores potential for much tighter coupling of cores
 - *Same deal for stream programming model!*
- Looking beyond SMP
 - Cache Coherency: *necessary but not sufficient*
 - Fine-grained language elements difficult to build on top of CC protocol
 - Hardware Support for Fine-grained hardware synchronization
 - Message Queues
 - Transactions: Protect against incorrect reasoning about concurrency



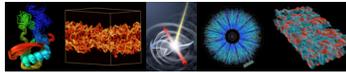
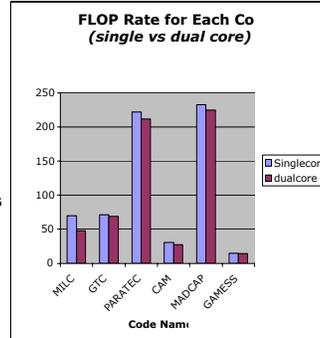
What about MPI?

- What about Message Passing on a chip?
 - MPI buffers & data structures growing $O(N)$ or $O(N^2)$ a problem for constrained memory
 - Redundant use of memory for shared variables and program image
- What about Message Passing on a million processor system?
 - Applications developers today write programs that are as complex as describing where every single bit must move between the 6,000 transistors of the 8080a.
 - We need to at *least* get to the “assembly language” level.
- We need to reconsider our entire programming model if this is indeed what the future holds for us.
 - My expectation: hybrid model with MPI and “something new” for the many core chips



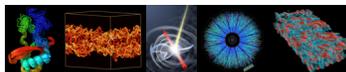
Will Multicore Slam Against the Memory Wall?

- Memory Bandwidth Starvation
 - *“Multicore puts us on the wrong side of the memory wall. Will CMP ultimately be asphyxiated by the memory wall?”* Thomas Sterling
 - Memory wall is NOT a problem that is caused by multicore (term coined in 1994).
- What about latency (other part of memory wall)
 - Effective use of bandwidth is progressively inhibited by poor latency tolerance of modern microprocessor cores (*memory mud rather than memory wall*)
 - Stalled clock rates actually halt growing gap of memory latency / operation
- We can fix bandwidth (but not latency)
 - With current technology, we could put 8x more bandwidth onto chips than we currently do! . . . GPUs and Cisco Metro already do this!
 - So why don't we do it? . . . because it is ineffective for current processor cores
 - Cell/Software controlled memory can use bandwidth more effectively



Mass Migration to New Algorithms

- Materials Science
 - Predict bulk material properties from first principles (ab-initio)
 - One algorithm, Planewave DFT, accounts for 75% of the materials science workload
 - Codes: QBox, PARATEC, VASP
 - QBox won Gordon Bell award for scalability!
- However, this is *not* the correct algorithm to use for petaflop scale calculations!
 - FLOP requirements grow $O(N^3)$
 - Increasingly dominated by BLAS3 (*good for FLOPs*)
 - But only get to simulate marginally larger system
 - Fails to exploit locality of quantum wave component!
- Classical DFT approach cannot continue!
 - $O(N)$ algorithms will eventually replace them
 - $O(N)$ methods are not yet fully developed because the attention is going to classical DFT because it generates impressive FLOP rates
 - 75% of the NERSC MatSci workload is going to have to migrate to $O(N)$ methods, but little support that migration



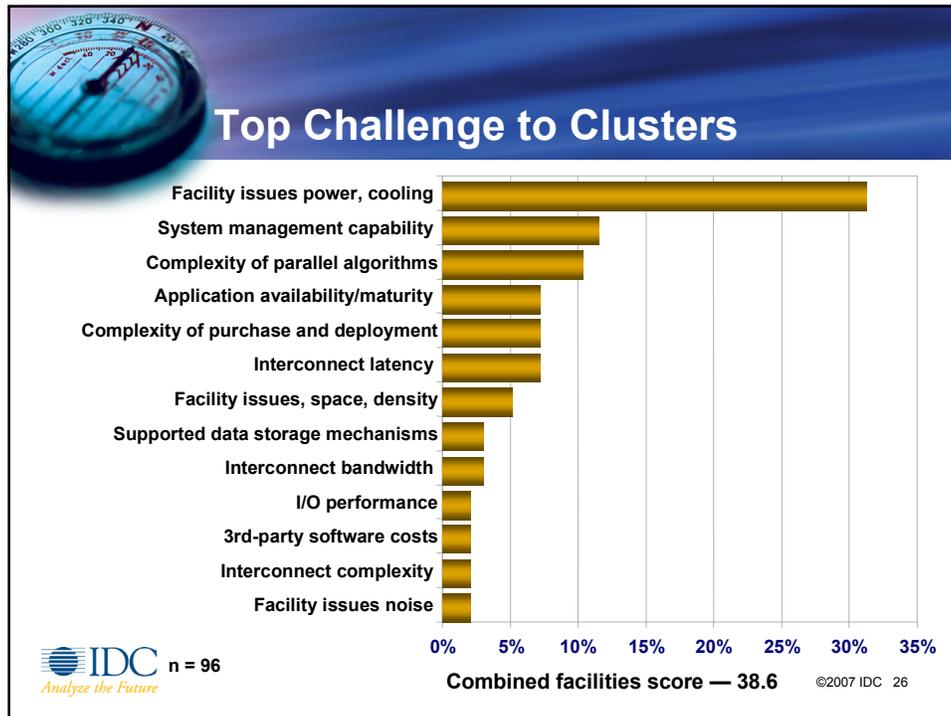
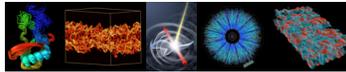
Overview

News from NERSC

My (TOP500) Predictions for Petascale Computing

Three Challenges for Petascale Computing:

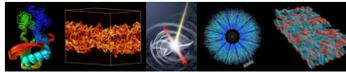
- Parallelism
- **Power**
- The Petascale bubble



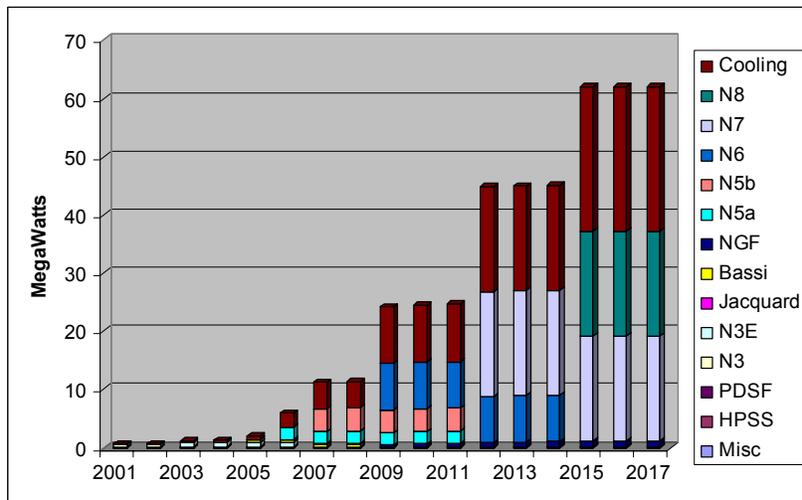
NERSC Estimate ...

... for a sustained Petaflops system (on multiple applications) in 2010

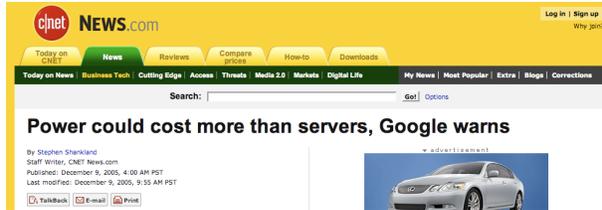
- 20 MW
- 16,000 square feet
- \$12M/year electricity cost



NERSC Projections for Computer Room Power System + Cooling



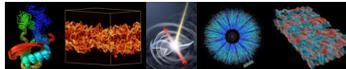
Power is an Industry Wide Problem



The New York Times "Hiding in Plain Sight, Google Seeks More Power",
by John Markoff, June 14, 2006



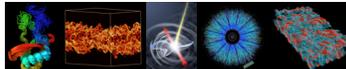
New Google Plant in The Dalles, Oregon,
from NYT, June 14, 2006



The transition to low power technology is inevitable

Does it make sense to build systems that require the electric power equivalent of an aluminum smelter?

- Information "factories" are only affordable for a few government labs and large commercial companies (Google, MSN, Yahoo ...)
 - Midrange installations will soon hit the 1 - 2 MW wall, requiring costly new installations
 - Economics will change if operating expenses of a server exceed acquisition cost
- The industry will switch to low power technology within 3 - 4 years
- Embedded processors or game processors will be the next step (BG, Cell, and SiCortex)



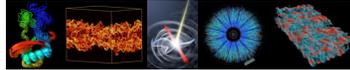
Increasing Blue Gene Impact

- **SC 2005 Gordon Bell Award, 101.7 TFs on real materials science simulation**
 - Recently exceeding 200 TFs sustained
- **Sweep of the all four HPC Challenge class 1 benchmarks**
 - G-HPL (259 Tflop/s), G-RandomAccess (35 GUPS) EP-STREAM (160 TB/s) and G-FFT (2.3 Tflop/s)
- **Over 80 large-scale applications ported and running on BG/L**



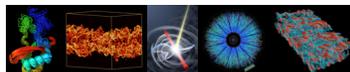
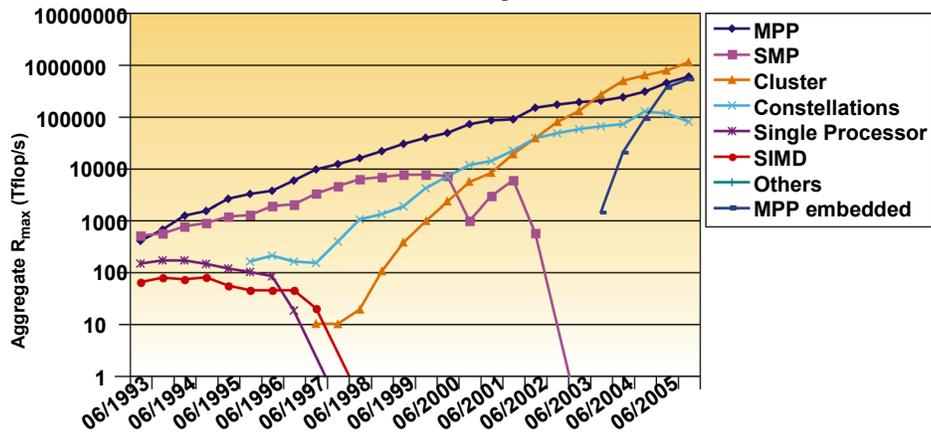
27.6 kW power consumption per rack (max)
7 kW power consumption (idle)

Slide adapted from Rick Stevens, ANL



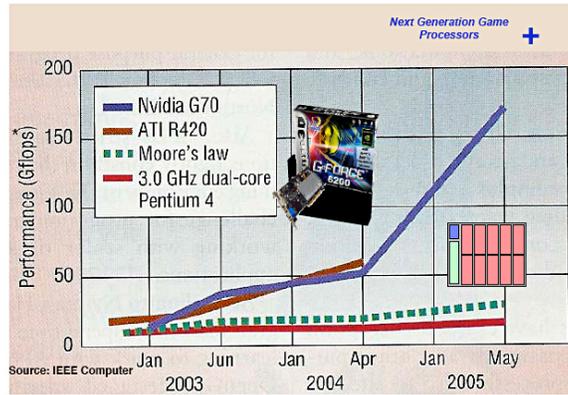
BG/L—the Rise of the Embedded Processor

TOP 500 Performance by Architecture



Increasing Game Processor Impact

- GPUs & Game Processor Architectures Are an Excellent Match for Game Applications
- Performance Has Been Growing Faster Than Moore's Law !

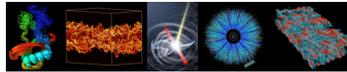


*Single Precision

• Typical Street Price for High End PC Graphics Card: **300 - 400 US\$**



Source: Randy Moulic, IBM

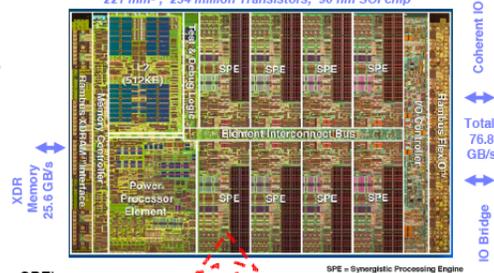


IBM Cell

Supercomputing Capability for High Volume, Consumer Systems

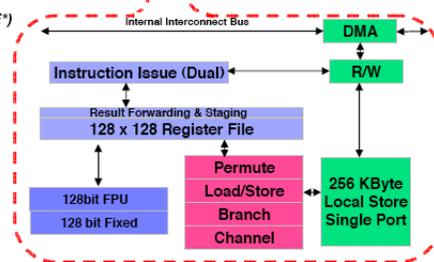
221 mm², 234 million Transistors, 90 nm SOI chip

- Multi-core, multi-thread, "cluster-on-a-chip"
 - 64bit PowerPC Control Processor
 - + 8 Tightly integrated accelerators (SPE)
 - 128 bit SIMD/Vector, MAC
 - 256KB Embedded Memory
 - Integrated I/O and memory interfaces
- High Performance
 - 3.2 GHz clock frequency
 - 205 GFLOPs/s peak, single precision (dual issue, in-order execution, 25.6 GFLOPs per SPE)
 - ~20 GFLOPs/s peak, double precision (2 DP instructions every 7 cycles, 1.83 GLOPs per SPE)
 - 205 GB/s internal interconnect bandwidth
 - ~100 GB/s BW for memory, external IO
- Linux OS
 - Simultaneous multiple OS support
 - + Real-time support



XDR Memory
25.6 GB/s

Coherent IO
Total: 76.8 GB/s
IO Bridge



Source: Randy Moulic, IBM

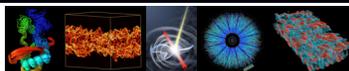
GPU's become general purpose

... for example Nvidia's CUDA

CUDA & GPU Computing 

- **CUDA is a completely new architecture and programming model for general-purpose computation on GPUs**
- Hardware and Software designed together
 - NOT a new driver for old GPU architectures
- Data-parallel computing with thousands of threads
- Parallel Data Cache helps increase arithmetic intensity for massive speedups
- **Program in C**
- BLAS and FFT libraries

© NVIDIA Corporation 2008



... and Intel's reaction

From a presentation by Doug Carmean, Intel

CUDA & GPU Computing 

- **CUDA is a completely new architecture and programming model for general-purpose computation on GPUs**
- Hardware and Software designed together
 - NOT a new driver for old GPU architectures
- Data-parallel computing with thousands of threads
- Parallel Data Cache helps increase arithmetic intensity for massive speedups
- **Program in C**
- BLAS and FFT libraries

© NVIDIA Corporation 2008

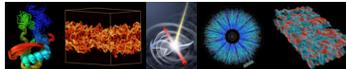
YIKES!

All dates, figures and product plans are preliminary and are subject to change without notice. Copyright © Intel Corporation 2008



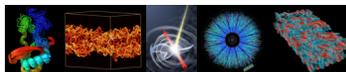
NERSC is addressing the power challenge at all levels

- Component level
 - Investigate use of low power components
- System level
 - Measuring and understanding energy consumption of system
 - Collaborate with EETD to improve energy efficiency
- Computer Room level
 - Understand airflow and cooling technology
- Building Level
 - Enforce rigorous energy standards in new computer building
 - Use of innovative energy savings technology



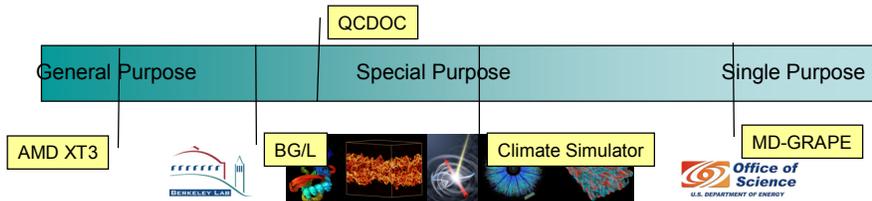
Accomplishments and Plans

- Component level
 - Evaluated cell processor for scientific calculations (30,000 downloads of paper) (see also new paper by Oliner et al. later in the conference)
 - Planning to work with Nvidia and ATI in 2007
- System level
 - Five year cost of ownership calculation for NERSC5 procurement
 - Developed conceptual plan for low power custom system for climate computing in collaboration with Tensilica
 - LDRD about measuring power consumption in progress
 - Planning to introduce new metric into TOP500
 - Energy standards of servers (J. Koomney, Stanford and LBNL)



Architectural Study of Climate Simulator (paper by J. Shalf, L. Oliker, and M. Wehner)

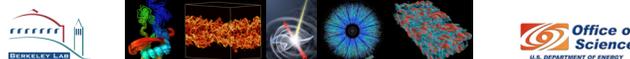
- We design system around the requirements of the km-scale climate code.
- Examined 3 different approaches
 - AMD Opteron: Commodity Approach - Lower efficiency for scientific applications offset by cost efficiencies of mass market
 - Popular building block for HPC, from commodity to tightly-coupled XT3.
 - Our AMD pricing is based on servers only without interconnect
 - BlueGene/L: Use generic embedded processor core and customize System on Chip (SoC) services around it to improve power efficiency for scientific applications
 - Power efficient approach, with high concurrency implementation
 - BG/L SOC includes logic for interconnect network
 - Tensilica: In addition to customizing the SOC, also customizes the CPU core for further power efficiency benefits but maintains programmability
 - Design includes custom chip, fabrication, raw hardware, and interconnect
- Continuum of architectural approaches to power-efficient scientific computing



Petascale Architectural Exploration

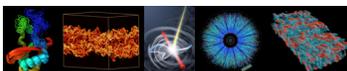
Processor	Clock	Peak/ Core (Gflops)	Cores/ Socket	Mem/ BW (GB/s)	Network BW (GB/s)	Sockets	Power (based on current generation technology)	Cost (based on current market price)
AMD Opteron	2.8GHz	5.6	2	6.4	4.5	890K	179 MW	\$1.8B
IBM BG/L	700MHz	2.8	2	5.5	2.2	1.8M	27 MW	\$2.6B
Climate computer	650MHz	2.7	32	51.2	34.5	120K	3 MW	\$75M

- AMD and BG/L based on list price
 - Of course discount pricing would apply, but extrapolation gives us baseline.
- Is it crazy to create a custom core design for scientific applications?
 - Yes, if the target is a small system.
 - In \$100M Petaflops system development costs are small compared to component costs.
 - In this regime, customization can be more power and cost effective than conventional systems.
 - Berkeley RAMP technology can be used to assess the design's effectiveness before it is built.
- Software challenges (at all levels) are a tremendous obstacle for any of these approaches.
 - Unprecedented levels of concurrency are required.
- This only gets us to 10 Petaflops *peak* - thus cost and power are likely to be 10x-20x more.
 - However, in ~5 years we can expect 8-16x improvement in power- and cost-efficiency.

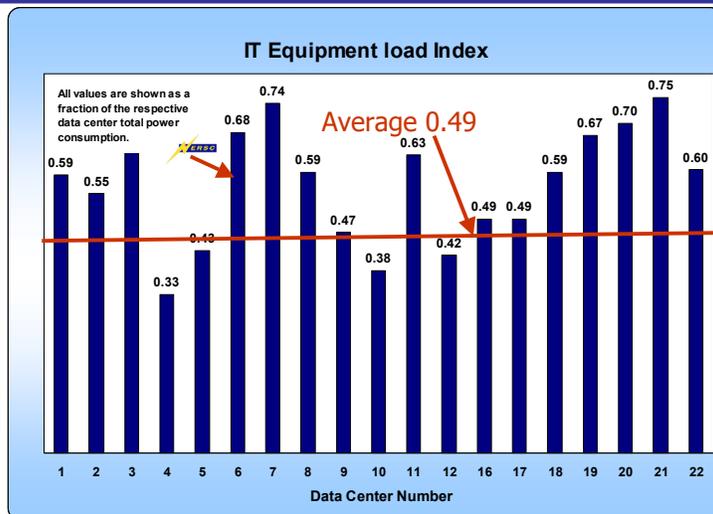


Working with the Experts in Energy Efficiency

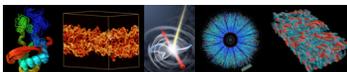
- Computer Room level
 - Measurements and modeling of computer room airflow in OSF
 - Worked with EETD on new measurement technology
- Building Level
 - Explored alternative building cooling (outside air)
 - Recirculation of hot computer room air
 - Discussion of energy savings technology with Chevron (others planned in 2007)



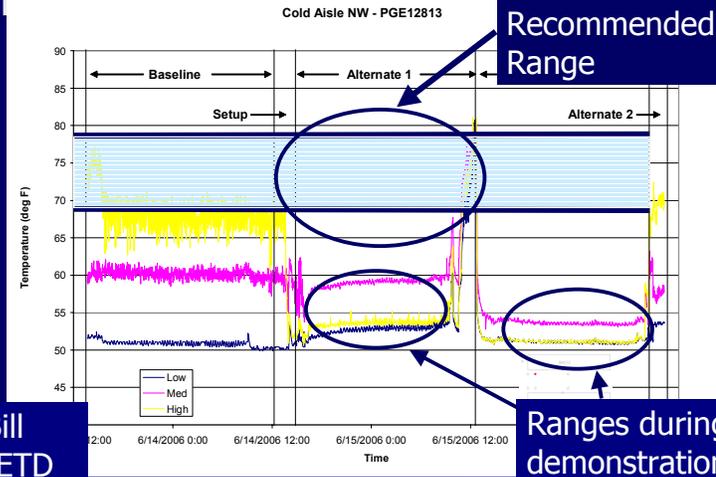
Fraction of Power to IT Equipment



Study by Bill Tschudi,
EETD, LBNL

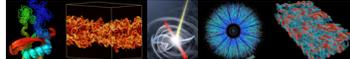
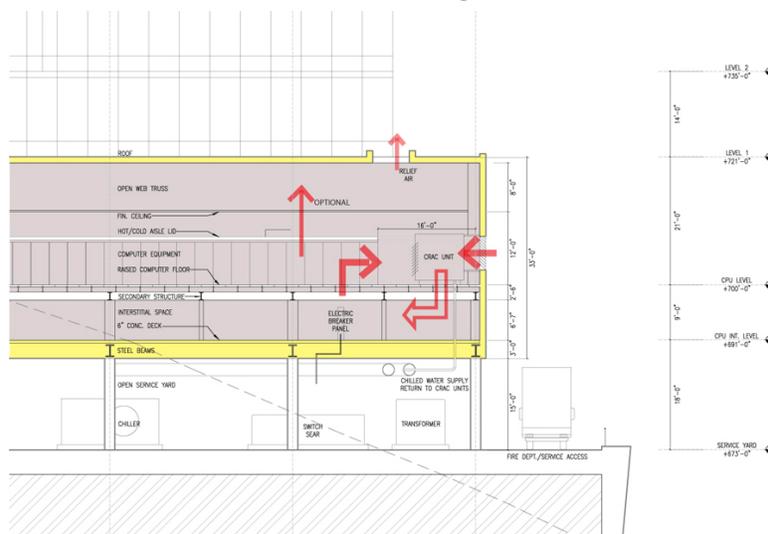


Better temperature control would allow raising the temperature in the entire data center

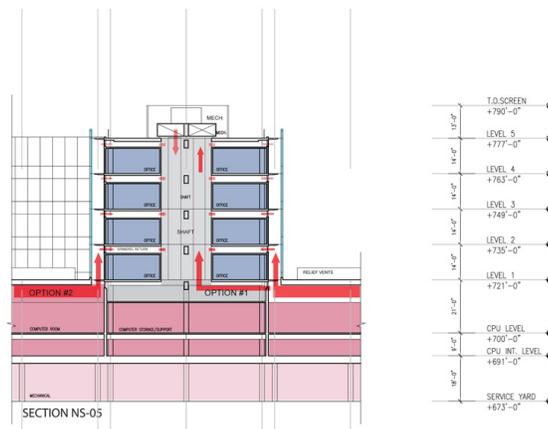


Study by Bill Tschudi, EETD

Plan for Use of Outside Air in New Building



Recirculation of Computer Room Hot Air



See High-tech Buildings website for latest publications:
<http://hightech.lbl.gov>



Overview

News from NERSC

My (TOP500) Predictions for Petascale Computing

Three Challenges for Petascale Computing:

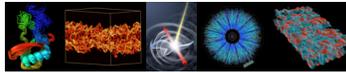
- Parallelism
- Power
- **The Petascale bubble**



2007 Petascale Mania

- Many sites made announcements in the last six months
 - Mostly still vaporware, delivery in the future
- National initiatives (Japan, Europe)
- DARPA HPCS phase 3
- NSF Petascale Track 1 competition

The reality is: even today there is only one general purpose system worldwide with >100 Tflop/s LINPACK performance



A Petaflops before its Time

- Even among experts there is an undue optimism about how close we are to “Petascale” computing
- In 11/2008 there will be a (Linpack Rmax) Petaflops computer on the TOP500 list
- Most likely it will be a BG/P, or a hybrid like the LANL Opteron/Cell “roadrunner”
- It will create an unwarranted sense of accomplishment
- It will distract from the development of real production Petaflops systems



Ushering in True Petascale Computing: Challenge and Opportunity 2007 - 2016

- **All of computing** will be highly parallel by 2010
- The current HPC ecosystem will radically change
 - Architectural innovation, even for processors
 - MPI + “something” programming model
- HPC centers will take systems approach to energy management
- Being on the forefront of these challenges, HPC has the opportunity to completely redefine computing

